

## Thesis abstract

# Data-efficient and generalizable machine learning in complex environments

Xiaobo Xia

Abstract of a thesis for a Doctorate of Philosophy submitted to  
the Faculty of Engineering, The University of Sydney, Australia

In an age marked by an unprecedented influx of data across diverse domains, the quest for effective machine learning (ML) solutions has increased significantly. However, data imperfections in complex environments present formidable obstacles, encompassing defective, redundant, and scarce data. Specifically, defective data, characterized by annotation errors and incompleteness, obstruct the learning process, particularly in critical domains such as healthcare and finance. Redundant data overwhelm relevant insights, demanding efficient filtering techniques for optimal ML performance. Besides, scarce data that are prevalent in domains with limited examples, necessitate robust ML models capable of generalizing effectively. Addressing these challenges is pivotal for unlocking the full potential of ML technologies. This thesis offers innovative solutions across three key areas: learning with defective data, redundant data, and scarce data. Particularly, for defective data, it explores learning with mislabelled and incomplete data, which proposes novel methods for handling each scenario. In the realm of redundant data, the thesis introduces a moderate coreset selection technique to enhance ML effi-

ciency across diverse practical tasks, and a refined coreset selection strategy to reduce the size of the constructed coreset while maintaining satisfactory model performance. Additionally, it addresses the challenge of scarce data by proposing advanced strategies for kernel mean estimation and augmenting datasets by marginalized corruption distributions to improve sample efficiency and model generalization. This thesis provides comprehensive insights and solutions for learning with imperfect data. By addressing these obstacles, it promotes the development of data-efficient and generalizable ML, and lays the groundwork for transformative breakthroughs in fields such as healthcare, finance, and climate science, propelling innovation and progress fuelled by the power of ML.

Dr Xiaobo Xia  
Postdoctoral Fellow  
National University of Singapore

E-mail: [xiaoboxia.uni@gmail.com](mailto:xiaoboxia.uni@gmail.com)

URL: <https://ses.library.usyd.edu.au/handle/2123/32939>